# Assignment-based Subjective Questions

1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?

   Answer :

- Multiple users have availed the bikes in summerand fall,and in 2019

- Around Aug , Sept and Octber there is seen a spike in the usage
- in Holiday people seems to have availed the bike very less and in not holiday situation as usual very high number of users are booking the bikes. As seen the median is quite higher than of a median of holiday.
- if workingday more number of users are availing the bikes
- Weather situation has a good trend for prediction as well. weathersit 1 and 2 is where most of the users has availed bikes

2. Why is it important to use drop_first=True during dummy variable creation?
   Answer:

   As an example there are three categories x , y and z .
   When we create dummy variable it shows like below:

| x | y | z |
|---|---|---|
| 1 | 0 | 0 |
| 0 | 1 | 0 |
| 0 | 0 | 1 |

   It is clearly visible where ever there is 1 that category value is true .However the question is do we really need 'x' . Because when y and z both as 0 it is definite that x will be 1 hence y and z is enough to represent x and this resolves multicollinearity as well.
   Although I have used OHE in the assignment.

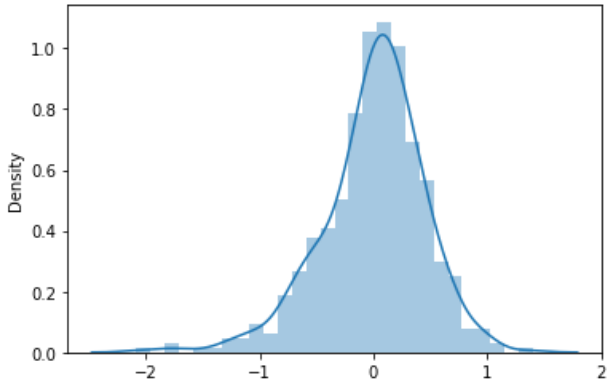
3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?

Answer:

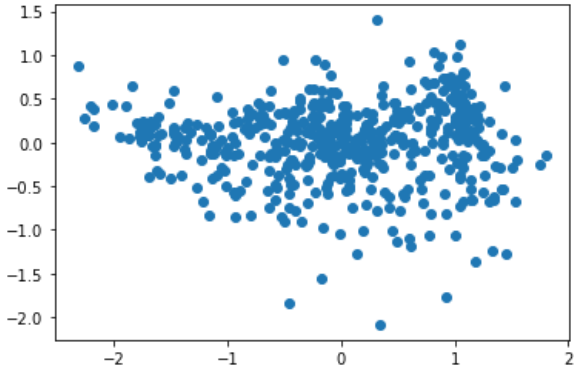
temp(Temperature)

4. How did you validate the assumptions of Linear Regression after building the model on the training set?

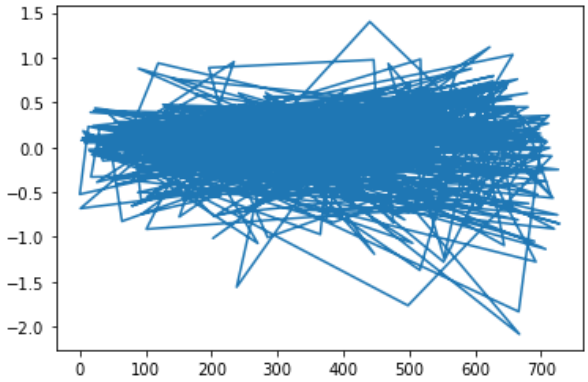
- I checked Normality of Error




- Checked Homoscedasticity




- Autocorrelations of Errors




5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?

Answer:

As per the final model - Weather Situaion_3 (Light Rain) , holiday and temp.

# General Subjective Questions

1.Explain the linear regression algorithm in detail.

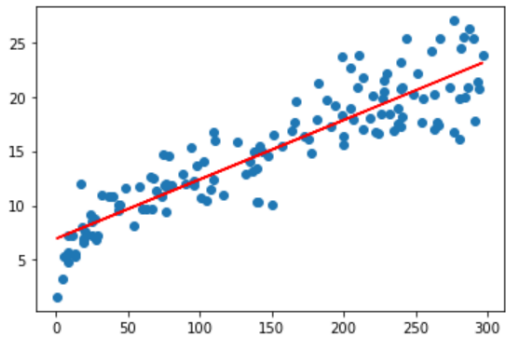
Answer: Linear regression is a supervised learning (knowing labels from the beginning) Build for continuous variables. Basically, it predicts the target variable.

As an example, if we provide a set of variables where there are variables to learn from and when an unknown value is given, based on the training or knowledge it gathered from the training/learning variables it predicts some answer for the unknown variable.

The main algorithm is based on y = mx + c formula for a straight line since it is a linear algorithm. Here y is the predicted value for each x provided. It builds a linear model which is a straight line based on the training data x and y values. And when a new x is given it gives us y value as per the straight line generated based on the training set.

```
# 6.9487+0.0545*X_train = lr_model.predict(X_train_sm)

plt.scatter(X_train,y_train)
plt.plot(X_train,y_train_pred,'r')
plt.show()
```



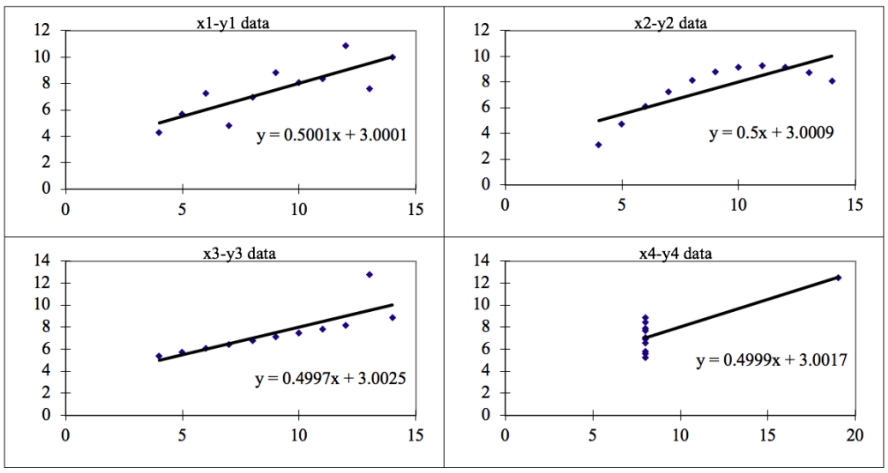

Here you see the blue dots made by training x and y. and the straight line was made with the help of x values fitting into y = b0 + b1x formula.

There are two kinds of linear regression simple (having 1 value of x to learn from), multiple when there are multiple values of x learn from, although all the x's might not help in that case we need to drop which is linearly dependent.

2. Explain the Anscombe's quartet in detail.

Answer:
Anscombe's Quartet can is a set of four identical but when plotted they appear differently. It is constructed by Francis Anscombe to makes understand that plotting graphs before creating a model is very important. This can help identify anomalies present in dataset.

Examples Outliers , diversity of data, linear separability etc.
The linear regression can only be built and fit for the data with a linear relationship.




The 1st Dataset describes the linear regression in a well manner.

The 2nd Dataset could not describe a linear relation or regression model as it's a non-linear data.

The 3rd Dataset describes the outliers can't be handled by linear regression.

The 4th Dataset describes the outliers can't be handled by linear regression.

All important features must be visualized before implementing any machine learning algorithm which helps in a good fit model.

3. What is Pearson's R?

Pearson's R, it is a measure of how good two variables are correlated between each other. It is a correlation coefficient looks for the quantitative value of how these variables are connected linearly.

This is denoted by 'r'.

A very good example would be age and height, humidity and rain, rainy season and umbrella sells.

Assumptions:

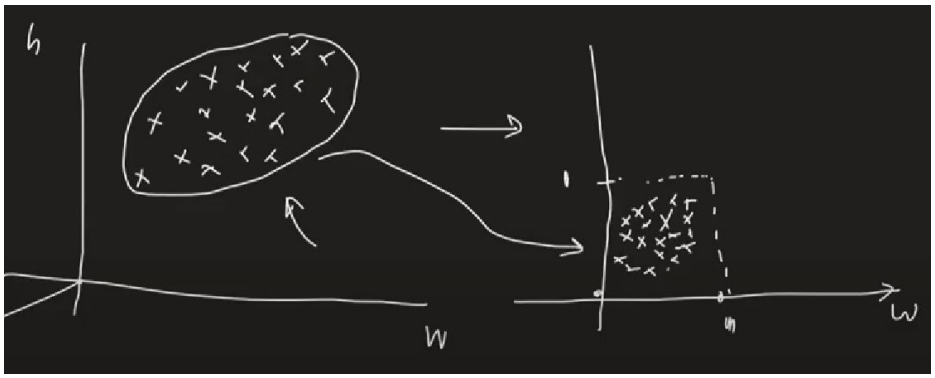
- Both variables should be normally distributed
- No significant outliers
- Has a linear relationship
- Homoscedascity
- Continuous variables

4.What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?

Answer:

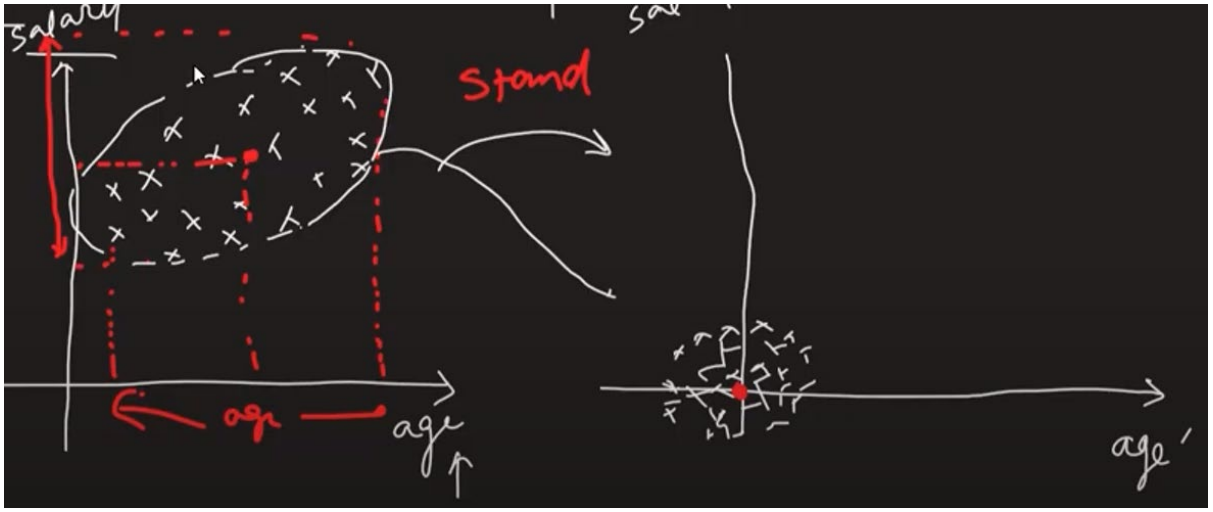
The datasets that has to be used for training or testing might have numerical data sets which might vary a lot in terms of magnitude or range. It impacts for the model when learning the coefficients as it can swing and give a not recommended value.

So normalize scaling as an example minmax scaling squeeze  the complete data between 0 and 1.




Xi^ = (Xi -Xmin)/Xmax - Xmin

And in standardization it is '0' cantered and standard deviation becomes 1.




Xi^ = (Xi -Xmean)/sigma

6. You might have observed that sometimes the value of VIF is infinite. Why does this happen?

   Answer:
   If there is a perfect correlation between the independent variables then Vif is infinite.
   R2 = 1/(1-R2)
   R2 is 1 for perfect correlation. Hence the r2 becomes inf (1/0)
7. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression

   Answer: A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second data set. By a quantile, we mean the fraction (or percent) of points below the given value.

Importance of Q-Q plot:

- Since Q-Q plot is like probability plot. So, while comparing two datasets the sample size need not to be equal.
- Since we need to normalize the dataset, so we don't need to care about the dimensions of values.

Reference: https://towardsdatascience.com